DOCUMENT RESUME

ED 480 129 TM 035 205

AUTHOR	Bashook, Philip G.
TITLE	Case Specificity Is Essential for Valid Performance Assessment of Physicians.
PUB DATE	2002-07-00
NOTE	10p.; Paper presented at the Ottawa Conference on Assessment (Ottawa, Ontario, Canada, July 13-16, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Case Studies; Medical Education; Medical Evaluation; *Performance Based Assessment; *Physicians; *Validity
IDENTIFIERS	*Specificity

ABSTRACT

Assessment of clinical performance is intended to make generalizations about a physician's competence. Performance assessments involve three interacting components: the candidate, the cases, and the raters. Confidence in generalizations depends on reliability of the measurements and whether the cases represent the competence domains of interest. Since clinical tasks and decisions are not general traits, but integral to specific patient cases, how many and what cases are sufficient to overcome case specificity. This paper offers a solution by adapting work by A. LaDuca to define an assessment blueprint and select representative cases that sample accurately the competence domain core and defined border (construct validity). Other concerns are measurement precision and score reproducibility (reliability) that are influenced by the quantity of cases and the way measurements are taken. Improved reliability requires controlling for variability across cases, raters, and exam administration using structural (e.g., balanced case selection, rater training) and statistical controls. (Contains 21 references.) (Author/SLD)

ABSTRACT

Case Specificity Is Essential For Valid Performance Assessment Of Physicians

Philip G. Bashook
University of Illinois at Chicago
Department of Medical Education

Keywords: performance assessment, construct validity, case specificity, score reproducibility, measurement precision, case sampling, reliability

Assessment of clinical performance is intended to make generalizations about a physician's competence. Performance assessments involve three interacting components, the candidate, the cases, and the raters. Confidence in generalizations depends upon reliability of the measurements and whether the cases represent the competence domains of interest. Since clinical tasks and decisions are not general traits but integral to specific patient cases, how many and what cases are sufficient to overcome case specificity? This paper offers a solution by adapting work by A. LaDuca to define an assessment blueprint and select representative cases that sample accurately the competence domain core and defined borders (construct validity). Other concerns are measurement precision and score reproducibility (reliability) that are influenced by the quantity of cases and the way measurements are taken. Improved reliability requires controlling for variability across cases, raters and exam administration using structural (e.g., balanced case selection, rater training) and statistical controls.

Paper presented at the
Ottawa Conference on Assessment, Ottawa, Canada, July 13-16, 2002
Contact: P Bashook at pgb@uic.edu

Case Specificity Is Essential For Valid High Stakes Performance Assessments*

Philip G. Bashook
University of Illinois at Chicago

Many psychometricians recommend applying the same psychometric analysis to performance assessments based upon traditional test theory (1) they employ for measuring declarative knowledge on standardized exams (2,3,4, 5). Adopting this recommendation-estimated exam reliabilities frequently are below 0.65 in a typical one-day performance exam with 10-12 authentic cases. (6,7,8). After adjusting reliability estimates for administrative, rater, and format error the remaining measurement error is attributed to "case specificity." (9 p 86) Case specificity is defined as the error variance from specific and unique characteristics of each authentic case or real-world situation in the exam.(10) The recommended solution is to add more cases (6,7,9 p.147). Lower exam reliabilities may be tolerable when ranking candidates to provide constructive feedback, but not when exam scores reward professionals with certificates or licenses. Adding more cases to improve exam reliability with equivalent cases or to broaden the case sample can double or triple exam time, an impractical and very expensive solution.

What to do? Maybe the problem is not that "case specificity" increases measurement error, but that traditional test theory has lead us astray. The real issue in my view is validity. How does the content-related evidence for items represent the content of a defined practice domain? And, what is the underlying construct for the test scores that purport to measure practical performance on authentic cases? (11,12). Following the

**P Bashook,* Case Specificity Is Essential For Valid High Stakes Performance Assessments, paper presented at Ottawa Conference, Ottawa, Ontario, Canada, July 13-16, 2002

reasoning about construct validity articulated by Messick (12) and adopted in the recent Testing Standards (13) reliability is just one bit of evidence for the validity argument when interpreting performance assessment scores. This paper challenges the accepted psychometric dogma about case specificity as measurement error and argues that case specificity is essential for valid high stakes performance assessments. The paper analyzes assumptions about validity underlying this measurement conundrum with examples drawn from certification and licensure of physicians.

False assumption 1: there is an infinite universe of performance task items that can be sampled independent of cases (content validity)

Traditional test theory assumes a near infinite universe of test items.(1,3) This assumption works well for testing declarative knowledge using MCQs because with a large item universe reliability estimates can treat random item sampling as equivalent to representative sampling. Also, knowledge tests assume knowledge chunks are equivalent across test items and scores item performance as 1 or 0. Unweighted item scores are combined to derive a test score later corrected for reliability using for example generalizability statistics. (10) Statistical checks for differential item functioning (DIF) provide a means to identify item bias' that violate the sampling assumptions.(14)

In performance assessment the task item universe is finite and limited by what cases are selected and what content domains the cases represent. Consider for example performance tasks for a case involving the first doctor visit for a patient when diabetes is suspected but not yet diagnosed. In case two the patient has diabetes and the doctor visit

is for follow-up treatment to change the prescribed insulin regimen. Case one can measure performance in making the diagnosis of diabetes, and case two can measure adjusting treatment. Ignoring these case differences when randomly sampling diabetic cases would not provide a content valid test score. Also, specific tasks in the first case are not the same as in the second, and task scores combined across cases produces a meaningless test score.

If one domain to be tested is diagnostic acumen then the first case might be selected from a small pool of cases that specifically require diagnosing only diabetic cases (a small item universe), or a larger item pool containing cases that concern only diagnostic challenges across many diseases. Case selection would not be random from either item pool, but cases carefully chosen to obtain a representative sample of case performance situations that comprise the domain of practice. Clearly, the case sampling becomes stratified into content domains and further subdivided into a matrix sampling problem when considering different medical specialties. The practice domain for internists who should have expertise in diabetic diagnosis and complex management is quite different from the surgeon who relies upon internists for advice in diabetic care during surgery. LaDuca calls this the "practice model."(15) When using the practice model the test blueprint for performance assessments must specify the expected depth and scope of professional practice for a professional discipline using exemplar authentic cases and tasks for each case. Cases excluded from the practice model are irrelevant for purposes of assessing performance. Selecting a random sample of cases drawn from insurance records, or hospital records would be influenced by the type of insurance coverage and case mix bias

in hospital admissions and hospital community demographics. The case sample might provide a useful database for building the practice model, but it is not a generalizable case universe.

Further compounding representative case sampling is the need to specify the stage in the evolving case situation (i.e., initial diagnosis, treatment planning; 16,17) and the precise tasks that must be performed to demonstrate case mastery. Authentic cases "require one to recognize a problem space; to plan strategies, to take initial steps, and gather additional information; and, observing preliminary results, to determine which direction to proceed." (18) For some cases all the aforementioned tasks need to be assessed, but typically a case is deconstructed into smaller components or performance tasks that are essential to successful case management. Page and Bordage refer to these essential judgments as "key features" of the case (19). Each case-based task generates a case specific performance score that can be combined into a case score similar to combining checklist ratings to score a standardized patient case (20). Case scores are the unit of measure when generating an exam score. The content validity of the exam score depends upon representative sampling of cases from the practice profile, not random case sampling or task sampling across distinctly different cases. It is no surprise that exam reliability estimates are low when the exam content validity is questionable.

False assumption 2: Performance scores measure an ability construct (construct validity)

The score on a measured case-based task cannot be divorced from the underlying case situation. The validity of the task score depends directly on relevance of the measured task to the case content. If the exam purpose is to measure competent performance (i.e., certification or licensure) than content validity also depends upon whether the identified and measured case specific tasks distinguish a competent from a less competent performer on the case. Conversely, by deconstructing a practical case into essential tasks and measuring those tasks, the score on the measured tasks for a case infers performance on the case not a hypothetical ability that generalizes across case situations. Even for basic skills the case situation regulates interpretation of task performance.

For example, surgeons consider suturing skin (stitch together a cut in the skin) a very basic skill that must be mastered and is performed with nearly every surgical procedure. Measurements of suturing performance might be done by observing suturing on real people, on a mannequin or plastic model of skin tissue, or with a virtual reality environment. To demonstrate this task there must be a specific person, mannequin or model with skin to suture, otherwise what is the performance? The exam authors decide about the problem space by determining the purpose of suturing (e.g., wound repair, trauma repair, retain opening for drainage tube), patient characteristics, skin texture (e.g., fat, damage), wound condition, and available suturing tools (e.g., different threads and needles, staples, plastic material). Change any of these variables and the performance tasks change. There is no underlying construct called suturing. The suturing tasks are unique to the individual case and differ as the case varies. Measurements of suturing

skills, like history taking or physical examination skills, do not generalize across cases. Not even problem solving ability can be generalized across cases (21).

Conclusion

Evidence of construct validity for performance measurement scores begins with content validity building upon a practice profile. The logical sequence for case specificity in valid performance measurement is:

1. Practice profile begets test blueprint,
2. Which defines case selection and case specific tasks,
3. Which controls what tasks must be measured for each case,
4. Which generates task scores per case that must be converted into case scores,
5. Which are combined to produce the case-based exam score that is valid for the test blueprint based upon the practice profile.

In summary, case specificity is essential for content and construct validity in performance assessment. Reliability estimates that attribute measurement error to case specificity or combine task scores independent of cases should be questioned for their veracity.

References cited

1. Nunnally, JC. *Psychometric Theory*. New York: McGraw-Hill Book Co.; 1967.
2. Case, SM, Swanson, DB. *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners; 1996.
3. Haladyna, TM. *Developing and Validating Multiple-choice Test Items*. Hillsdale, New Jersey: L. Erlbaum Associates; 1994.

**P Bashook,* Case Specificity Is Essential For Valid High Stakes Performance Assessments, paper presented at Ottawa Conference, Ottawa, Ontario, Canada, July 13-16, 2002

4. Brennan, RL. An Essay on the History and Future of Reliability from the Perspective of Replications. *Journal of Educational Measurement.* 2001; 38(4): 295-317.

5. Linn, RL, Burton, E. Performance-Based Assessment: Implications of Task Specificity. *Educational Measurement: Issues and Practice.* 1994(Spring 1994):6-15.

6. van der Vleuten, CP. The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education.* 1996;1:41-67.

7. Swanson, DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, Eds. *Further Developments in Assessing Clinical Competence.* Montreal, Canada: Canada Health Publishers; 1987:13-45.

8. Swanson, DB, Norman, GR, Linn, RL. Performance-Based Assessment: Lessons From the Health Professions. *Educational Researcher.* June/July 1995 1995;24(5):5-11, 35.

9. Elstein, AS, Shulman, LS, Sprafka, SA. *Medical Problem Solving: An Analysis of Clinical Reasoning.* Cambridge, Massachusetts: Harvard University Press; 1978.

10. Shavelson, RJ, Baxter, GP, Gao, X. Sampling Variability of Performance Assessments. *Journal of Educational Measurement.* Fall 1993 1993;30(3):215-232.

11. Cronback, LJ. Five Perspectives on Validity Argument. In: Wainer H, Braum HI, Eds. *Test Validity.* Hillsdale, New Jersey: L. Erlbaum Assoc.; 1988:3-17.

12. Messick, S. The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In: Wainer H, Braum HI, Eds. *Test Validity.* Hillsdale, New Jersey: L. Erlbaum Assoc.; 1988:33-45.

13. Joint Committee, AERA, APA, NCME. *Standards for Educational and Psychological Measurement.* Washington, DC: American Educational Research Association; 1999.

14. Holland, PW, Wainer, H. *Differential Item Functioning.* Hillsdale, NJ: L. Erlbaum Assoc.; 1993.

15. LaDuca, A. Validation of Professional Licensure Examinations. *Evaluation & the Health Professions.* June 1994 1994;17(2):178-197.

P Bashook, Case Specificity Is Essential For Valid High Stakes Performance Assessments, paper presented at Ottawa Conference, Ottawa, Ontario, Canada, July 13-16, 2002

16. Shavelson, RJ, Baxter, GP, Pine, J. Performance Assessments: Political Rhetoric and Measurement Reality. *Educational Researcher.* 1992;May 1992(-):22-27.

17. Bashook, PG, Hruska, L. A Conceptual Framework and Authoring Tool for Constructing Case-based Simulations. Paper presented at American Educational Research Association; April 2003; Chicago.

18. Mislevy, RJ. Foundations for a new test theory. In: Frederiksen N, Mislevy RJ, Bejar II, Eds. *Test Theory for a new Generation of Tests.* Hillsdale, New Jersey: L. Erlbaum Assoc; 1993:19-39.

19. Page, G, Bordage, G, Allen, T. Developing Key-Feature Problems and Examinations to Assess Clinical Decision-Making Skills. *Academic Medicine.* Mar 1995;70(3):194-201.

20. van der Vleuten, C, Swanson, DB. Assessment of Clinical Skills With Standardized Patients: State of the Art. *Teaching & Learning in Medicine.* 1990;2(2):58-76.

21. Page, GG. Assessing reasoning and judgment. In: Mancall EL, Bashook PG, Eds. *Assessing clinical reasoning: the oral examination and alternative methods.* Evanston, IL: American Board of Medical Specialties; 1995:19-27.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035205

I. DOCUMENT IDENTIFICATION:

Title: Case Specificity is Essential for Valid Performance Assessment of Physicians

Author(s): Philip G. Bashook

Corporate Source: University of Illinois

Publication Date: July 13-16, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Philip G Bashook

Printed Name/Position/Title: Philip Bashook, Res. Asst. Prof

Organization/Address: University of Illinois at Chicago, Dept Medical Education (MC591), 808 S. Wood St, Chicago IL 60612

Telephone: 312 996-5448

FAX: 312-413-2048

E-Mail Address: PGB@uic.edu

Date: 8/15/2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions